Towards a chemistry-informed paradigm for designing molecules

Srinivas Rangarajan

Department of Chemical & Biomolecular Engineering, Lehigh University, Bethlehem PA 18015

srr516@lehigh.edu

+1 610-758-4219

Abstract

Computational design of molecules for optimal performance is of interest in many fields, including chemical engineering. Often, however, these methods, in particular those based on rigorous mathematical optimization, do not explicitly take into consideration chemistry information, such as (but not limited to) synthesis feasibility. This opinion article discusses traditional and current approaches through examples from the literature where properties that depend on chemical transformations of the molecule are incorporated in the design process. Through these examples, the article highlights the importance of cheminformatics, graph theory, and machine learning in: (1) representation of the molecules, (2) reaction prediction and generation, and (3) property estimation. The article finally presents a vision of including information about chemical transformations in molecule design procedures, highlighting rigorous optimization and machine learning approaches such as generative modeling and reinforcement learning.

I. Introduction

Design of molecules for a chemical process is of tremendous interest to chemical engineers and chemists. The sheer size of the molecule universe, often estimated[1] to be on the order of $10^{20\text{-}60}$, implies that experimental approaches are likely insufficient to explore this space comprehensively. Computational approaches broadly termed computer-aided molecule design (CAMD)[2-4], are therefore more tractable in discovering promising molecular candidates for a target application. Such approaches have been pursued for decades and include: (1) explicit enumeration of molecules (either queried from a database such as PubChem or constructed from fragments) and subsequent screening based on a chosen property metric[5]; (2) evolutionary optimization based techniques to search the space using expert heuristics[6, 7]; (3) rigorous mathematical optimization including derivative-free optimization, to construct molecules from building blocks such that one or more properties are optimized[8-11]; and (4) more recently, machine learning and artificial intelligence based approaches that include generative modeling to sample molecules[12], continuous optimization over a reduced-dimensional learned latent space[13], and reinforcement learning to dynamically optimize the molecular structure to maximize rewards (e.g. a chosen property metric)[14, 15]. Furthermore, the process systems engineering (PSE) community has correctly identified the multiscale nature of molecule design by integrating this step with process-level information[16-18].

These approaches enable identifying a molecule (or a set of molecules) that can be further examined experimentally, thereby rendering the problem of molecule design substantially more tractable, however, they are often not cognizant of the underlying chemistry-specific constraints. Chemistry information can be pertinent to molecule design for at least three reasons. First, molecules identified by the chosen approach should be synthetically feasible (or synthesizeable), i.e., they should be synthesized using available raw materials (feedstocks) and well-established and selective chemistries. Mathematically rigorous methods such as optimization-based design or artificial intelligence may identify molecules with superior properties than the reference (or benchmark) molecules; however, these methods also often identify unrealistic structures that may be too energetically unstable or require several synthesis steps to be produced cost-effectively in an industrial setting[19]. Second, in addition to requiring synthetic feasibility, plausible chemistry-based restrictions may arise from a sustainability standpoint. For instance, there may be a desire to produce the molecule from renewable sources and/or using benign chemistries. Concepts such as "bioprivileged" [20] molecules, i.e., whether or not a molecule can be made from biomass[21], and the popular metrics such as E-factor[22] and atom economy, that can be used to determine the greenness of a molecule synthesis procedure, require ascertaining the synthesis routes to make the target molecule from specific set of building blocks and chemistries. Third, in several cases such as the design of fuels, energy carriers, solvents, and molecular catalysts, the underlying process chemistry may be integral to the performance of the molecule and cannot be easily delinked. Specific examples include: (1) being aware of the charging/discharging chemistry while designing liquid organic hydrogen carriers (LOHCs)[23]; (2) tracking the decomposition chemistry to determine the environmental impact of molecules[24-26]; and (3) taking into account how intermediates and transition states of the reaction network interact with the molecular catalyst or solvent[27]. Finally, rather than designing a molecule for a specific target, chemistry information is also relevant in charting and analyzing the synthesis landscape for product portfolio design. Specifically, mapping out the whole network of synthesis options available, starting from a given set of reactants and using known chemistries, provides a wholistic (or "systems") understanding of “what can be made?” and “how can they be made?”. For instance, one could address questions such as: *what are the molecules that can be made from lignin monomers using heterogeneous catalysis? Which of them are feasible candidates as fuels*?

Incorporating chemistry information, beyond synthetic feasibility, in the overall process of designing molecules is well-recognized. While there are specific examples in the literature that accomplish this, there are no systematic and rigorous frameworks that have been pursued to integrate molecular characteristics related to underlying chemical transformations within the problem of molecule design. Even synthesis feasibility is arguably often treated as a secondary requirement with primacy given to molecular performance metrics. This opinion article focuses on current chemistry-informed approaches to molecule design. The next section introduces general concepts useful in molecule design and tracking chemistry while the subsequent section discusses specific examples from the literature where chemistry information has been incorporated at different stages of the design process. This opinion will skip the detailed discussion of property-based molecule design methods in view of other comprehensive works on that topic. Finally, this article ends with an outlook for seamlessly imbuing molecule design procedures with chemistry cognizance.

II. General concepts

Chemical graph theory: Fundamental to automated processing of chemical information is the representation of molecules as graphs, wherein atoms represent the nodes and the bonds represent the edges[28] (Figure 1). The nodes and edges are further annotated with atom-specific and bond-specific information such as atom type, charge, unpaired electron count, bond order, etc. Consequently, many concepts from algorithmic graph theory, such as breadth- and depth-first graph traversal, automorphism and subgraph isomorphism, cycle detection, etc., and from spectral graph theory, such as graph Laplacian, adjacency matrix, eigenvalues and eigenspectrum, etc., have been employed to represent, characterize, and manipulate molecular structures. More recently, graph kernels[29] and graph convolutional networks[30] find place in the context of machine learning for molecules.

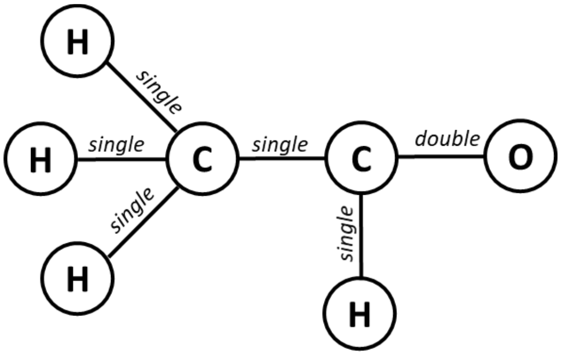


Figure 1: Graph theoretic representation of acetaldehyde. Atom and bond attributes are included.

Molecular property prediction: CAMD, per se, requires the prediction of relevant properties using a reliable method. While ab initio methods including quantum chemistry and molecular simulations offer robust predictive framework to compute electronic to bulk properties, the associated computational cost often render them as intractable options. Data-driven methods, trained on experimental or computational data (or both) offer a more tractable solution particularly if a large number of property evaluations are necessary for identifying suitable candidates. Traditionally, such models included group additivity[31] (Figure 2a) and quantitative structure property relations[32] while more recently the use of machine learning, ranging from generalized sparse additive models[33, 34] to deep neural networks[35-37], has gained significant attention due to data availability in many cases and advances in computational methods and theories to represent discrete data structures such as graphs. In particular, graph convolutional neural networks to identify latent spaces as well as predict molecular properties[38, 39] (Figure 2b) have gained significant attention. In cases where data is limited or hard to acquire, the emerging ideas include transfer[40] and multitask learning[41] whereby correlated data-rich and data-lean properties (tasks) can be learned together (sequentially as in the case of transfer learning and concurrently in the case of multitask learning). Further, a judicious design of the training set using concepts such as active learning[42] can also minimize the cost of building reliable data-driven models.

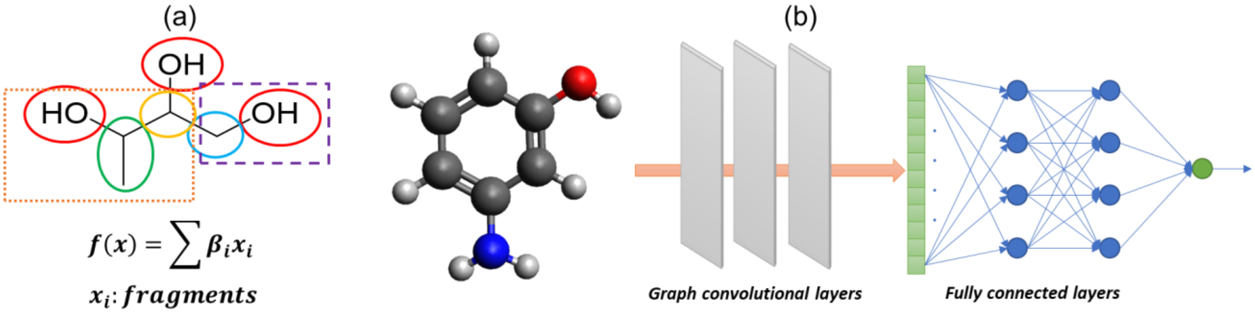


Figure 2 (a) An illustrative schematic of group contribution: Butane-1,2,3-triol is decomposed into first (solid), second (dashed), and third (dotted) order groups (the formula for group contribution is also shown); (b) A schematic of a graph convolutional neural network to compute properties from molecular structure as input. The group contribution formula shown in (a) is an easy-to-interpret linear model involving the summation of a regressed contribution, β, corresponding to each group (fragment) i, times the occurrence of the fragment, x_i.

Generation of synthesis routes: Synthesis pathways can be generated either in the forward direction, i.e., starting from the reactants and proceeding forward until target molecules are reached, or in the reverse direction (retrosynthesis) wherein the formation of a given molecule is traced backwards until easily available molecules are reached. In both approaches, the traditional way to generate reactions and intermediates is through the use of expert-determined reaction templates (or reaction rules) as shown in Figure 3. These templates contain information about the plausible set of chemical transformations that can occur; for instance, double bond hydrogenation, Friedel-Crafts alkylation, and the Suzuki coupling can all be considered as generic reaction rules. When these templates are applied to the molecular graph of the reactant (or the product in the case of retrosynthesis), new graphs are generated corresponding to the product (or the reactant in the case of retrosynthesis), thereby leading to the generation of new reactions. The templates can be applied iteratively to generate a sequence of steps that relate the initial reactant and the final product. Computer-aided synthesis planning (CASP), or computer-aided organic synthesis (CAOS), has largely revolved around retrosynthesis[43, 44] because the goal is to produce a specific molecule (the target drug, for instance). On the other hand, forward synthesis is more relevant when the synthesis landscape needs to be explored or pathways to a class of molecules (e.g., alcohols) needs to be identified.

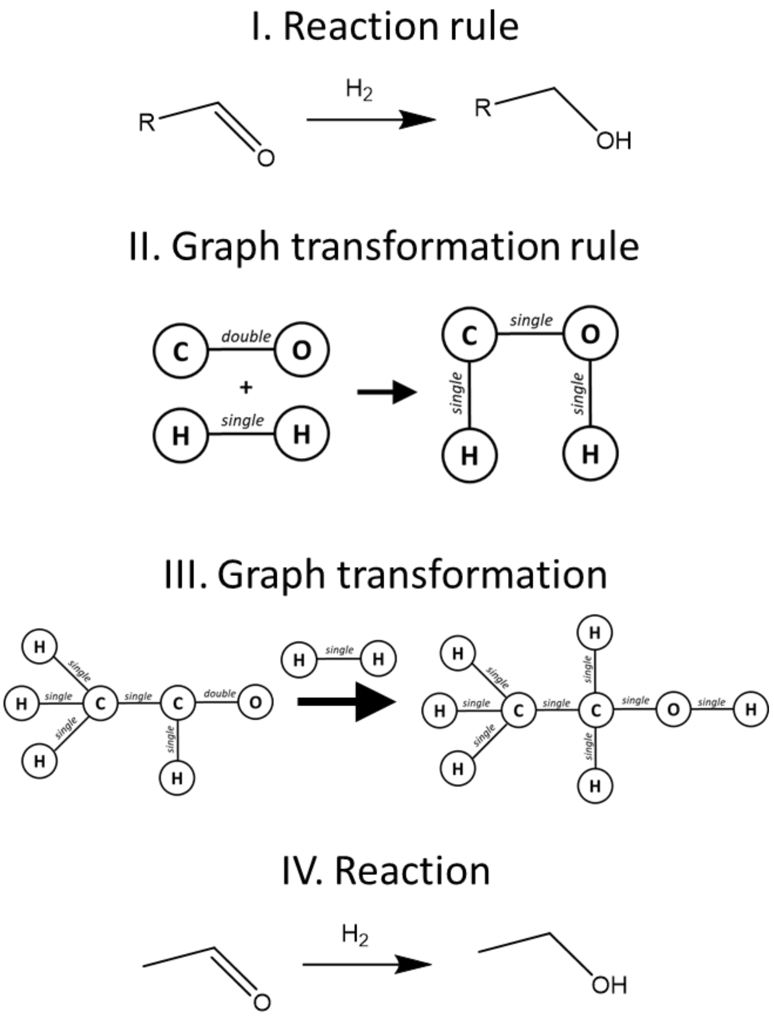


Figure 3: A schematic of graph-based generation of reactions using reaction templates (rules). I. shows a carbonyl hydrogenation rule; II. shows the graph representation of this rule; III. shows the application of graph transformation rule to the reactant graph (Figure 1) to produce product graph (ethanol); IV. shows the resulting hydrogenation reaction.

Forward synthesis is usually carried out using rule-based reaction network generators[45]. A recent example of such tools is Rule Input Network Generator (RING) developed by Rangarajan et al[46-48]. As shown in Figure 4, RING accepts as input initial reactants and reaction rules, written in the form of a program in a domain-specific chemistry specification language. These instructions are then used by a network generator that iteratively applies the rules to the reactants and products generated thereof to construct a comprehensive reaction network that is complete and correct with respect to the inputs. RING also accepts instructions and queries to extract information from the reaction network, such as identifying pathways connecting the reactants with the products.

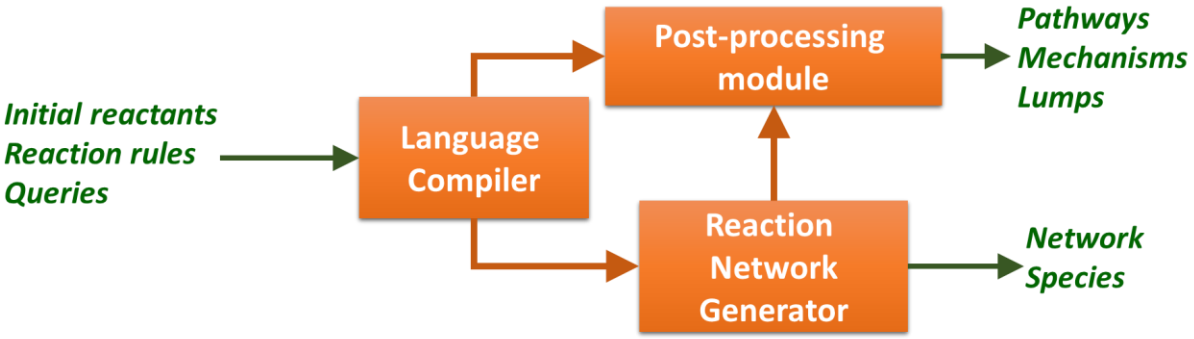


Figure 4: Schematic of inputs, outputs, and components of RING, an example of a rule-based network generator for forward generation of synthesis network and routes.

III. Approaches and examples of chemistry cognizant molecule selection

This section sketches a few approaches that directly or indirectly use chemistry information while designing molecules. Most of the applications in the literature focus on synthesizeability (or synthesis routes) either during or after molecule design; applications involving chemistry information other than synthesis (forward or reverse) are relatively scarce but are also discussed to highlight the different types of chemistry information that may need to be tracked.

Synthesizeability check as a postprocessing step: A straight-forward way to account for synthesizeability in design approaches is to use it as a secondary screening criterion once molecules have been designed/selected based on performance criteria. Such an evaluation can be via: (1) explicit identification of the retrosynthetic pathway so that potential bottlenecks can be determined a priori, or (2) computation of synthesis or molecular complexity scores as relatively inexpensive surrogate indicators of how easy it is to synthesize the target molecule. Retrosynthesis tools (CASP) can be used to identify the synthesis routes in the first method; as mentioned earlier, traditionally these tools utilized reaction templates and expert heuristics to identify promising pathways[49]; more recently, machine learned tools for reaction prediction have been effectively used for the same[50, 51]. For the second method, either traditional additive, fragment-based models can be used as surrogate scores for ease of synthesis and molecular complexity[52-55]. In particular, these models were usually trained on sets of molecules whose ease of synthesis was determined by organic synthetic chemists, whether or not they are available commercially or present in databases such as PubChem[56]. More recently, machine learned models have been used to evaluate synthetic feasibility. For instance, in one study, a neural network was trained on a reaction corpus to compute the synthesis scores of molecules with the constraint that the score of the product in a reaction in the corpus is greater than or equal to that of the reactants[57]. This ensures that the score is roughly correlated to the number of steps required to produce the molecule. In another study, a retrosynthesis tool was used on a collection of molecules to determine whether or not synthetic pathways could be found; subsequently, a machine learned classifier was trained to predict the result (i.e., success or failure) of the retrosynthetic tool[58]. While both approaches offer a means to incorporate synthetic feasibility into the overall workflow of molecule design, the drawback is that it is plausible that all of the solutions from the molecule design step, or at least a large fraction of them, may be deemed synthetically infeasible; from a computational cost standpoint, such a workflow would be inefficient.

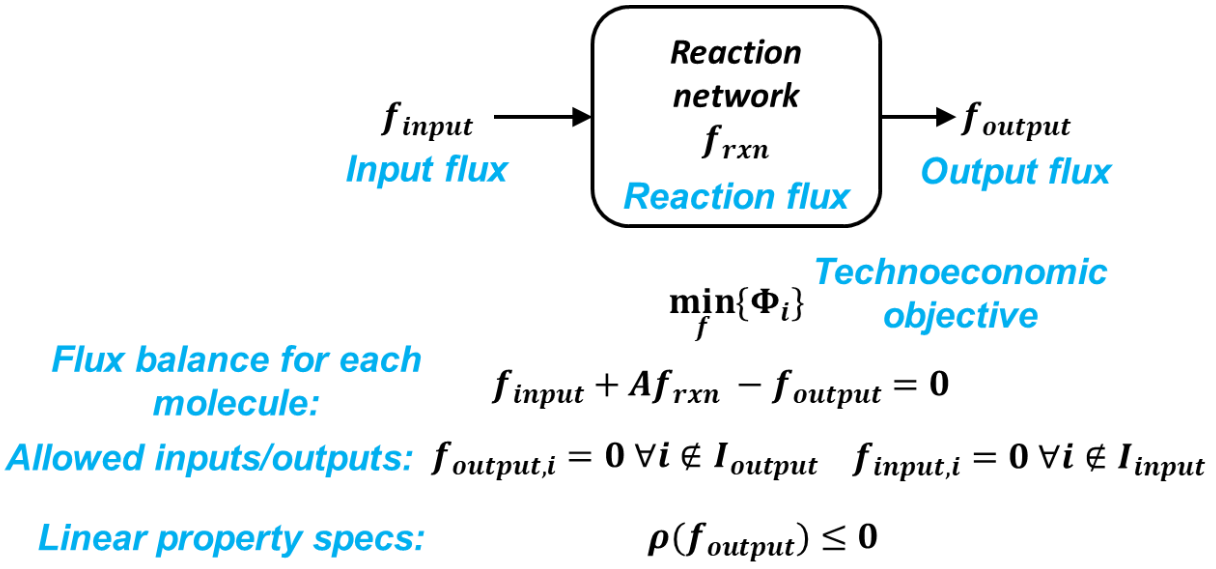
Incorporating synthetic feasibility checks during molecule screening: High-throughput virtual screening, i.e., the approach of explicit enumeration of large number of molecules and their subsequent evaluation one-by-one, can be modified so that the library of molecules that is screened is generated in a focused manner. For instance, well-documented synthesis reaction templates[59] can be systematically applied to easily available starting materials (and their products) to generate a large number of synthetically feasible molecules[60-62]. More recently, it has been shown that generative models[63, 64], e.g., using recurrent neural networks or autoencoders, can be used to sample new molecules that are similar in properties to a set of known molecules; these methods can then be employed to identify more synthesizeable molecules if the known molecules themselves are easily synthesized. Generation and evaluation of focused libraries are particularly popular in drug design to create lead libraries for virtual screening; however, they do not guarantee solutions with performance as good as optimization-based approaches.

Reaction network generators can also be used to generate these focused molecular libraries; the inputs (reactants and reaction rules) to these tools can be selected in such a way that the molecules generated by the network (or a specific subset of it) will constitute the library; if the reaction rules are derived from known chemistries, the molecules are in principle synthesizeable. For instance, Broadbelt and coworkers recently showed that, using biomass-derived molecules as reactants and common catalytic steps as reaction rules, network generation using their in-house software NETGEN can identify biopriviledged compounds[65]. Several of the generated compounds were already in databases such as PubChem database indicating that many of the industrially useful molecules have alternative pathways that may be more sustainable. Similarly, Rangarajan et al. used RING to generate the space of fatty alcohols that can be derived from biomass-based platform chemicals and known catalysis rules[66]. Such an approach could enable applying complex chemical characteristics (or information, in general), e.g., being biomass-derived or derived from biochemical routes, as constraints in molecule design.

Network flux analysis to optimally select synthesis routes and molecules: A number of molecules may satisfy property specifications for an application, multiple routes may exist to make each molecule, and a given raw material can be plausibly used to make many of the desired molecules. One therefore encounters two questions while designing molecules: *what to make? and how?* Reaction network flux analysis (RNFA), proposed by Marquardt and coworkers[67, 68], is quite valuable in addressing this. A reaction network of synthesis options comprising a set of potential products is first assembled in any suitable way (as will be further discussed). A flux-balance based optimization problem is then solved to select initial reactants, reactions (and thereby one or more pathways), and end products that satisfy user-specified criteria and objective. The problem can be set up as a linear program with the variables being material flow for each molecule and fluxes for each reaction as generically shown in Scheme 1; binary variables can be added (to determine if a reaction is selected or not) to prevent reaction cycles and find alternative solutions. The objective is usually a pathway-based technoeconomic metric, such as: (1) minimizing a cost function that depends on the selected reactions and their fluxes or (2) minimizing the largest reaction enthalpy barrier in the selected pathway. The flux balance shown in scheme 1 is a constraint to ensure that there is an uninterrupted pathway between initial reactants and end product. One or more products can be selected by the problem, e.g., a specific product or a blend of multiple products satisfies boiling point or heating value requirements).

Scheme 1. A simplified problem formulation for simultaneous product and synthesis route selection. $\boldsymbol{\Phi}$ *is the objective of choice and is a function of ‘*$\boldsymbol{f}$*, the molecular flows (input, output) for each compound and the flux through each reaction.* $\boldsymbol{A}$ *is the stoichiometry matrix of the synthesis reaction network and* $\boldsymbol{I}$

represents a set of compounds. The problem is formulated to identify pathways to one or more output compounds. The waste flux is set to be defined


f_{input}
Input flux
Reaction network
f_{rxn}
Reaction flux
f_{output}
Output flux
$\min_f\{\Phi_i\}$
Technoeconomic objective
Flux balance for each molecule:
$f_{input} + Af_{rxn} - f_{output} = 0$
Allowed inputs/outputs:
$f_{output,i} = 0\ \forall i \notin I_{output}$
$f_{input,i} = 0\ \forall i \notin I_{input}$
Linear property specs:
$\rho(f_{output}) \leq 0$

In the original work, the reaction network was manually assembled from the literature; Marvin et al.[69], on the other hand, used RING to generate a more comprehensive network of biomass-derived molecules and corresponding reactions. In particular, as schematized in Figure 5, carefully curated set of reaction rules based on proven heterogeneous catalytic transformations and the top DOE-determined biomass-based platform chemicals were input into RING to generate a network of several thousand reactions and species (which, in principle, are also bioavailable). This network contained several compounds that could be blended with gasoline. The network was then fed into a mixed-integer linear program that identified product portfolio (i.e., mixture of molecules) that could be blended with representative gasoline samples such that all of ASTM fuel standards could be satisfied. Concurrently, the synthesis routes to make each of the molecules from biomass were also identified. Multiple optimal solutions can be identified and analyzed further in terms of thermochemistry or process considerations (such as reaction or phase coupling) in a post-processing step[70]. The network, in principle, can also include a comprehensive set of reported reactions in the literature, e.g., the reaxys database. Lakpin and coworkers used this database to identify strategic molecules (via graph theory and network traversal, not RNFA) that could play a role in upcycling waste streams[71]. Arguably, one could apply RNFA to such a network although techniques such as reinforcement learning[72] can also be pursued in view of the sheer size of the network (and, consequently, of the optimization problem in RNFA). Finally, the network of synthesis options can also be generated using retrosynthesis pathways for a collection of target molecules[73].

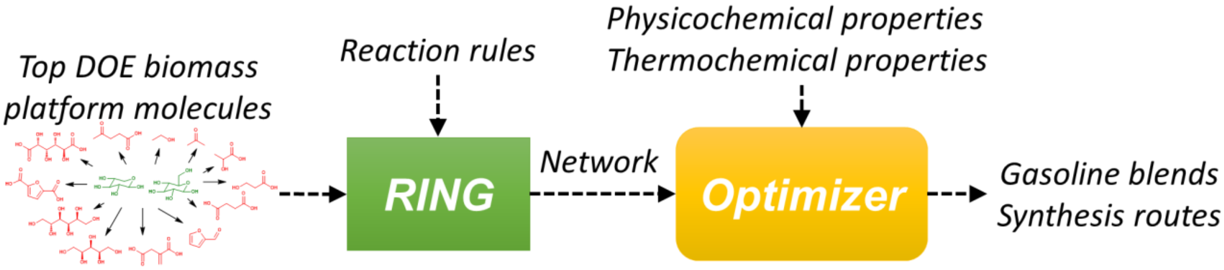

Top DOE biomass platform molecules
Reaction rules
RING
Network
Physicochemical properties
Thermochemical properties
Optimizer
Gasoline blends
Synthesis routes

Figure 5: Automated workflow for product and synthesis route selection using network generation and optimization.

The RNFA approach may also not yield molecules with performance on par with optimization-based approaches; however, it allows for effectively combining chemistry information and molecule selection into a single-step optimization problem. This provides a more systems viewpoint in design; a molecule with exquisite properties but that is not easily synthesizeable is, from a wholistic view, inferior to a synthesizeable molecule that has relatively poor performance but satisfies minimum property specs. Furthermore, the formulation allows for multiple objectives to be explored so that contrasting solutions may be identified. For instance, sustainability metrics (measured as the flux of waste CO_2 generated as byproduct, the E-factor of the products, or LCA specs[74]) can be introduced as an alternative objective to cost or thermochemistry and pareto-optimal solutions can be identified and analyzed. Synthesis network can also be controlled by precisely choosing which reactants and reaction rules are allowed; for instance, RING can be used to (1) generate molecules derivable from shale gas or carbon dioxide (analogous to bioprivileged) or (2) create synthons using electrochemistry, which can then be used as input to the flux balance analysis.

Incorporating chemistry characteristics beyond synthesizeability: Reaction network generation can also be used to explicitly include chemistry information beyond synthesis into the process of molecule selection. Consider the example of designing two-way liquid organic hydrogen carriers (LOHCs)[23]. Here, a hydrogen-lean molecule, such as toluene, is hydrogenated at the energy source using molecular hydrogen; the hydrogen-rich product, methylcyclohexane, the hydrogen carrier (i.e., the LOHC), can then be stored and transported to the point of energy demand where it can be dehydrogenated to produce hydrogen (Figure 6); the toluene molecule thus generated is recycled back to the source. One important property of LOHCs is the hydrogen storage capacity, i.e., the amount of hydrogen released per gram of the hydrogen-rich form. Determining this theoretical capacity automatically for any molecule (as is needed for molecule design), however, is nontrivial because to quantify the amount of hydrogen released by a single molecule, the hydrogen-rich and hydrogen-lean pair connected via a series of dehydrogenation reactions has to be identified. Furthermore, the practical storage capacity depends

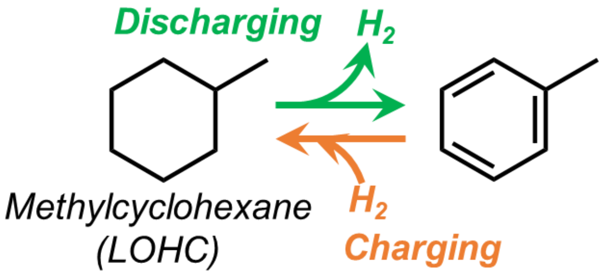


Figure 6: A schematic of two-way LOHC with discharging and charging chemistries.

on the kinetics and thermochemistry of the individual dehydrogenation steps (of which there can be multiple, each with varying thermokinetics). Paragian et al.[23] showed that the hydrogen-rich/lean pair and the dehydrogenation pathway(s) connecting the molecules can be determined using RING. In particular, a seed molecule from a molecule database (e.g., the Pubchem or the GDB database) can be taken and its fully hydrogenated and dehydrogenated forms can be identified using RING; these two molecules will form the LOHC pair and the reaction network will contain the pathways connecting them.

Network generation could be applied in a similar way to determine any molecular property that is dependent on some underlying chemical transformations. For instance, determining the environmental impact of a molecule may require identifying decomposition pathways and determining the toxicity or the degradation rates of the intermediates involved; a network generator can enumerate the decomposition reactions and extract the pathways[25].

IV. Vision: Towards a multiscale chemistry-cognizant molecule design

The examples discussed so far clearly indicate that incorporating chemistry information in molecule design is non-trivial and there may not be single optimization formulation that can be employed for all problems. However, it can be argued that that with machine learning, availability of data, and advanced optimization, chemistry information can be well-integrated with the problem of molecule design. This section lists a couple of directions as envisioned by the author.

Explicitly evaluating synthesis feasibility, bioprivilege of a molecule, or even the hydrogen capacity of LOHCs via forward generation or retrosynthesis within rigorous optimization-based approaches is computationally challenging. However, if surrogate metrics of these properties can be computed, they could be included as constraints within the optimization problem. As discussed earlier, such data-driven surrogates are already available for evaluating synthesizeability; similar scores can be derived for other chemistry-based properties of interest. For instance, the hydrogen capacity of a molecule can in principle be determined from its structure using a neural network model that is trained on a large set of hydrogen-rich/lean pairs generated using RING. Similarly, data-driven models can be developed to determine whether a molecule is bioprivileged or ease of separation of major and minor products in a synthesis sequence[75]. Given the nonlinearity of these properties with respect to structural information, however, these surrogates are likely to be highly nonlinear, thus rendering the optimization problem challenging to solve. However, the PSE and the larger optimization community has recently tackled such problems, for instance in the context of process synthesis and design[76-80]. Data-driven surrogate functions of more complex characteristics can also be included within such optimization problems. As an example, solutions of RNFA to produce specific molecules based on techoeconomic objectives can be used to train machine learned models that offer a more process-related metric than synthesis scores. Formulating such a network optimization problem (within RNFA) though will require surrogate models relating a reaction with a process metric such as capital or operating cost. Voll and Marquardt[67], and later Marvin et al.[69], used an empirical relation for cost based on energy loss across the process (which is roughly related to the reaction enthalpy) to construct the optimization objective. To be broadly applicable, however, more accurate and reliable data-driven approach that takes into consideration a wider array of processes and design parameters, such as models being trained on a large number of solutions of detailed and optimal design of various chemical processes, is required.

Alternatively, concepts from reinforcement learning can be combined with forward synthesis to generate synthetically feasible optimal molecules[81]. For instance, rather than create the entire network of synthesis options and then identify optimal molecules (and routes) within them, the network generation can be systematically biased to generate molecules with desired properties. Typical network generation process is comprehensive. However, to generate optimal molecules, not every intermediate needs to be processed further and not every rule needs to be applied to every molecule; only some of the intermediates and some of the rules (applied on specific types of intermediates) may lead to the generation of desired molecules. One could then bias the generation to only focus on these intermediates and rules by treating the problem as a Markov decision process (MDP) and adopt reinforcement learning techniques to solve it. Biased generation could directly lead to optimal molecules, or could instead be used to form a much smaller (and focused) network of synthesis options and RNFA can subsequently be applied to find optimal chemistries and molecules. Other complex chemistry-based metrics (e.g., hydrogen capacity of LOHCs) can also be incorporated by modifying the reward of the MDP appropriately.

Generative models that sample from a latent space could also be modified to incorporate chemistry information[82]. For instance, the training set for the underlying deep neural networks can be biased by only including synthetically feasible molecules. Alternatively, generative models can be trained for the whole reaction pathways so that sampling new pathways (and the associated product) will always ensure synthetic feasibility. Such techniques could be employed to other characteristics such as whether or not they can be generated from biomass.

Two final remarks end this section. First, one can note that the organic material space extends beyond single "0D" molecular space. Indeed, 1D chains (e.g. polymers), 2D sheets, and 3D molecular structures (e.g. covalent organic frameworks) can also be designed and synthesized usually by first constructing the building blocks (organic monomers) that can then polymerize to form larger, more complex structures. The need for chemistry informed molecule design approaches also arises in this context; for instance, while the space of COFs is large, not all of them are synthetically feasible or bioprivileged and incorporating such information while designing the building blocks requires the tools and approaches discussed above. Second, like any other molecular property, determining synthesizeability or any other property related to chemical transformations has an associated prediction error and uncertainty. It is important, therefore, to also consider these while quantifying the reliability of the solutions identified using a chosen design approach.

V. Conclusions

This opinion article discusses the concept of including chemistry information, including but beyond synthetic feasibility, while designing organic molecules. Multiple approaches have been considered in the literature, primarily in the context of synthesis, that explicitly (forward synthesis or retrosynthesis) or implicitly (synthetic feasibility scores) incorporate chemistry information during or after molecular screening or performance-based optimization. However, no comprehensive framework exists for chemistry cognizant molecule design. This article envisions that a combination cheminformatics and graph theory, optimization, and machine learning (including reinforcement learning) can provide this framework.

Acknowledgements: SR acknowledges support from Lehigh University. This work was partially supported by the National Science Foundation (Award # 1953245).

The author declares no conflict of interest.

References

1. Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A., Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of Computer-Aided Molecular Design* **2013,** *27* (8), 675-679.
2. Austin, N. D.; Sahinidis, N. V.; Trahan, D. W., Computer-aided molecular design: An introduction and review of tools, applications, and solution techniques. *Chemical Engineering Research & Design* **2016,** *116*, 2-26.
3. Ng, K. M.; Gani, R., Chemical product design: Advances in and proposed directions for research and teaching. *Computers & Chemical Engineering* **2019,** *126*, 147-156.
4. Gertig, C.; Leonhard, K.; Bardow, A., Computer-aided molecular and processes design based on quantum chemistry: current status and future prospects. *Current Opinion in Chemical Engineering* **2020,** *27*, 89-97.

• A discussion of how quantum chemistry methods, including COSMO approaches, can be used in molecule design, in particular, design of solvents and molecular catalysts

5. Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A., What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annual Review of Materials Research* **2015,** *45* (1), 195-216.
6. Sundaram, A.; Ghosh, P.; Caruthers, J. M.; Venkatasubramanian, V., Design of fuel additives using neural networks and evolutionary algorithms. *Aiche Journal* **2001,** *47* (6), 1387-1406.
7. Henault, E. S.; Rasmussen, M. H.; Jensen, J. H., Chemical space exploration: how genetic algorithms find the needle in the haystack. 2020; Vol. 2, p e11.
8. Samudra, A. P.; Sahinidis, N. V., Optimization-Based Framework for Computer-Aided Molecular Design. *Aiche Journal* **2013,** *59* (10), 3686-3701.
9. Conte, E.; Gani, R.; Ng, K. M., Design of Formulated Products: A Systematic Methodology. *Aiche Journal* **2011,** *57* (9), 2431-2449.

10. Liu, Q. L.; Zhang, L.; Liu, L. L.; Du, J.; Tula, A. K.; Eden, M.; Gani, R., OptCAMD: An optimization-based framework and tool for molecular and mixture product design. *Computers & Chemical Engineering* **2019,** *124*, 285-301.

• Presents a rigorous and generic MINLP formulation and solution for product and mixture design

11. Sun, Y. J.; Sahinidis, N. V.; Sundaram, A.; Cheon, M. S., Derivative-free optimization for chemical product design. *Current Opinion in Chemical Engineering* **2020,** *27*, 98-106.

• Presents methods and opportunities to employ derivative free methods to solve the challenging optimization problems in molecule design where first principle property prediction is required.

12. Sanchez-Lengeling, B.; Aspuru-Guzik, A., Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018,** *361* (6400), 360-365.

13. Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A., Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *Acs Central Science* **2018,** *4* (2), 268-276.

14. Zhou, Z. P.; Kearnes, S.; Li, L.; Zare, R. N.; Riley, P., Optimization of Molecules via Deep Reinforcement Learning. *Scientific Reports* **2019,** *9*.

15. Popova, M.; Isayev, O.; Tropsha, A., Deep reinforcement learning for de novo drug design. *Science Advances* **2018,** *4* (7).

16. Adjiman, C. S.; Galindo, A.; Jackson, G., Molecules matter: The expanding envelope of process design. *Computer Aided Chemical Engineering* **2014,** *34*, 55-64.

17. Schilling, J.; Tillmanns, D.; Lampe, M.; Hopp, M.; Gross, J.; Bardow, A., From molecules to dollars: integrating molecular design into thermo-economic process design using consistent thermodynamic modeling. *Molecular Systems Design & Engineering* **2017,** *2* (3), 301-320.

18. Adjiman, C. S.; Sahinidis, N. V.; Vlachos, D. G.; Bakshi, B.; Maravelias, C. T.; Georgakis, C., Process Systems Engineering Perspective on the Design of Materials and Molecules. *Industrial & Engineering Chemistry Research* **2021,** *60* (14), 5194-5206.

• Provides a PSE thinking to problems involving molecule and material (including catalyst) design

19. Gao, W. H.; Coley, C. W., The Synthesizability of Molecules Proposed by Generative Models. *Journal of Chemical Information and Modeling* **2020,** *60* (12), 5714-5723.

•• Presents how recent AI based generative models may not result in synthesizeable molecules and discusses strategies to improve synthesizeability in generative AI models.

20. Shanks, B. H.; Keeling, P. L., Bioprivileged molecules: creating value from biomass. *Green Chemistry* **2017,** *19* (14), 3177-3185.

21. Moity, L.; Molinier, V.; Benazzouz, A.; Barone, R.; Marion, P.; Aubry, J.-M., In silico design of bio-based commodity chemicals: application to itaconic acid based solvents. *Green Chemistry* **2014,** *16* (1), 146-160.

22. Sheldon, R. A., The E factor 25 years on: the rise of green chemistry and sustainability. *Green Chemistry* **2017,** *19* (1), 18-43.

23. Paragian, K.; Li, B. W.; Massino, M.; Rangarajan, S., A computational workflow to discover novel liquid organic hydrogen carriers and their dehydrogenation routes. *Molecular Systems Design & Engineering* **2020,** *5* (10), 1658-1670.

•• Proposes and demonstrates a workflow that is able to identify promising LOHC pairs that are connected by a sequence of dehydrogenation steps.

24. Mayeno, A. N.; Yang, R. S. H.; Reisfeld, B., Biochemical reaction network modeling: Predicting metabolism of organic chemical mixtures. *Environmental Science & Technology* **2005,** *39* (14), 5363-5371.
25. Wei, C. Y.; Rogers, W. J.; Mannan, M. S., Application of runaway reaction mechanism generation to predict and control reactive hazards. *Computers & Chemical Engineering* **2007,** *31* (3), 121-126.
26. Finley, S. D.; Broadbelt, L. J.; Hatzimanikatis, V., Computational Framework for Predictive Biodegradation. *Biotechnology and Bioengineering* **2009,** *104* (6), 1086-1097.
27. Struebing, H.; Ganase, Z.; Karamertzanis, P. G.; Siougkrou, E.; Haycock, P.; Piccione, P. M.; Armstrong, A.; Galindo, A.; Adjiman, C. S., Computer-aided molecular design of solvents for accelerated reaction kinetics. *Nature Chemistry* **2013,** *5* (11), 952-957.
28. Trinajstic, N., *Chemical graph theory*. Routledge: 2018.
29. Ghosh, S.; Das, N.; Goncalves, T.; Quaresma, P.; Kundu, M., The journey of graph kernels through two decades. *Computer Science Review* **2018,** *27*, 88-111.
30. Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P., Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design* **2016,** *30* (8), 595-608.
31. Hukkerikar, A. S.; Sarup, B.; Ten Kate, A.; Abildskov, J.; Sin, G.; Gani, R., Group-contribution(+) (GC(+)) based estimation of properties of pure components: Improved property estimation and uncertainty analysis. *Fluid Phase Equilibria* **2012,** *321*, 25-43.
32. Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A., Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chemical Reviews* **2010,** *110* (10), 5714-5789.
33. Li, B.; Rangarajan, S., Designing compact training sets for data-driven molecular property prediction. *Molecular Systems Design & Engineering*, 2019, 4(5), 1048-1057.

• Discusses how to design training sets for data-driven molecular property prediction

34. Gu, G. H.; Plechac, P.; Vlachos, D. G., Thermochemistry of gas-phase and surface species via LASSO-assisted subgraph selection. *Reaction Chemistry & Engineering* **2018,** *10.1039/C7RE00210F.*
35. Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T., A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies* **2020**.
36. Alshehri, A. S.; Gani, R.; You, F. Q., Deep learning and knowledge-based methods for computer-aided molecular design-toward a unified approach: State-of-the-art and future directions. *Computers & Chemical Engineering* **2020,** *141*.

• Presents a detailed comparative analysis of knowledge-based (traditional) and machine learning methods (modern) for molecule design.

37. Liu, Q. L.; Zhang, L.; Tang, K.; Liu, L. L.; Du, J.; Meng, Q. W.; Gani, R., Machine learning-based atom contribution method for the prediction of surface charge density profiles and solvent design. *Aiche Journal* **2021,** *67* (2).
38. Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. In *Neural message passing for quantum chemistry*, International Conference on Machine Learning, PMLR: 2017; pp 1263-1272.
39. Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P., Convolutional networks on graphs for learning molecular fingerprints. *arXiv preprint arXiv:1509.09292* **2015**.
40. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. In *A Survey on Deep Transfer Learning*, Cham, Springer International Publishing: Cham, 2018; pp 270-279.

41. Ruder, S., An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* **2017**.
42. Sener, O.; Savarese, S., Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489* **2017**.
43. Todd, M. H., Computer-aided organic synthesis. *Chemical Society Reviews* **2005,** *34* (3), 247-266.
44. Szymkuć, S.; Gajewska Ewa, P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski Bartosz, A., Computer-Assisted Synthetic Planning: The End of the Beginning. *Angewandte Chemie International Edition* **2016,** *55* (20), 5904-5937.
45. Broadbelt, L. J.; Pfaendtner, J., Lexicography of kinetic modeling of complex reaction networks. *Aiche Journal* **2005,** *51* (8), 2112-2121.
46. Rangarajan, S.; Bhan, A.; Daoutidis, P., Language-oriented rule-based reaction network generation and analysis: Applications of RING. *Computers & Chemical Engineering* **2012,** *46*, 141-152.
47. Rangarajan, S.; Bhan, A.; Daoutidis, P., Language-oriented rule-based reaction network generation and analysis: Description of RING. *Computers & Chemical Engineering* **2012,** *45*, 114-123.
48. Rangarajan, S.; Bhan, A.; Daoutidis, P., Rule-Based Generation of Thermochemical Routes to Biomass Conversion. *Industrial & Engineering Chemistry Research* **2010,** *49* (21), 10459-10470.
49. Hoffmann, R. W., Computer-Aided Synthesis Planning. In *Elements of Synthesis Planning*, Hoffmann, R. W., Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, 2009; pp 145-148.
50. Coley, C. W.; Green, W. H.; Jensen, K. F., Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research* **2018,** *51* (5), 1281-1289.
51. Segler, M. H. S.; Preuss, M.; Waller, M. P., Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018,** *555*, 604.
52. Boda, K.; Seidel, T.; Gasteiger, J., Structure and reaction based evaluation of synthetic accessibility. *Journal of Computer-Aided Molecular Design* **2007,** *21* (6), 311-325.
53. Ertl, P.; Schuffenhauer, A., Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* **2009,** *1*.
54. Bottcher, T., An Additive Definition of Molecular Complexity. *Journal of Chemical Information and Modeling* **2016,** *56* (3), 462-470.
55. Li, J.; Eastgate, M. D., Current complexity: a tool for assessing the complexity of organic molecules. *Organic & Biomolecular Chemistry* **2015,** *13* (26), 7164-7176.
56. Kim, S.; Chen, J.; Cheng, T. J.; Gindulyte, A.; He, J.; He, S. Q.; Li, Q. L.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E., PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research* **2021,** *49* (D1), D1388-D1395.
57. Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F., SCScore: Synthetic Complexity Learned from a Reaction Corpus. *Journal of Chemical Information and Modeling* **2018,** *58* (2), 252-261.

58. Thakkar, A.; Chadimova, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J. L., Retrosynthetic accessibility score (RAscore) - rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chemical Science* **2021,** *12* (9), 3339-3349.

• This work develops an AI classifier to predict the synthesizeability of a molecule trained on retrosynthetic pathways predicted by a synthesis planning tool.

59. Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S., A Collection of Robust Organic Synthesis Reactions for In Silico Molecule Design. *Journal of Chemical Information and Modeling* **2011,** *51* (12), 3093-3098.
60. Chevillard, F.; Kolb, P., SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. *Journal of Chemical Information and Modeling* **2015,** *55* (9), 1824-1835.

61. Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B., AllChem: Generating, searching, and manipulating 1020 synthetically accessible structures. *Abstracts of Papers of the American Chemical Society* **2007,** *233*, 238-238.

62. Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J. B., The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *Journal of Chemical Information and Modeling* **2016,** *56* (7), 1253-1266.

63. Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P., Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *Acs Central Science* **2018,** *4* (1), 120-131.

64. You, J.; Liu, B.; Ying, R.; Pande, V.; Leskovec, J., Graph convolutional policy network for goal-directed molecular graph generation. *arXiv preprint arXiv:1806.02473* **2018**.

65. Zhou, X. W.; Brentzel, Z. J.; Kraus, G. A.; Keeling, P. L.; Dumesic, J. A.; Shanks, B. H.; Broadbelt, L. J., Computational Framework for the Identification of Bioprivileged Molecules. *Acs Sustainable Chemistry & Engineering* **2019,** *7* (2), 2414-2428.

•• This paper discusses a workflow that uses a network generator, NETGEN, to identify molecules that can be made from easily available biomass-derived platform molecules and well-established chemistries.

66. Rangarajan, S.; Bhan, A.; Daoutidis, P., Identification and analysis of synthesis routes in complex catalytic reaction networks for biomass upgrading. *Applied Catalysis B-Environmental* **2014,** *145*, 149-160.

67. Voll, A.; Marquardt, W., Reaction network flux analysis: Optimization-based evaluation of reaction pathways for biorenewables processing. *AIChE Journal* **2011,** *58* (6), 1788-1801.

68. Dahmen, M.; Marquardt, W., Model-Based Formulation of Biofuel Blends by Simultaneous Product and Pathway Design. *Energy & Fuels* **2017,** *31* (4), 4096-4121.

69. Marvin, W. A.; Rangarajan, S.; Daoutidis, P., Automated Generation and Optimal Selection of Biofuel-Gasoline Blends and Their Synthesis Routes. *Energy & Fuels* **2013,** *27* (6), 3585-3594.

70. Allan, D.; Marvin, W. A.; Rangarajan, S.; Daoutidis, P., Optimization and Analysis of Chemical Synthesis Routes for the Production of Biofuels. In *Computer Aided Chemical Engineering*, Elsevier: 2015; Vol. 37, pp 1103-1108.

71. Weber, J. M.; Lio, P.; Lapkin, A. A., Identification of strategic molecules for future circular supply chains using large reaction networks. *Reaction Chemistry & Engineering* **2019,** *4* (11), 1969-1981.

72. Khan, A.; Lapkin, A., Searching for optimal process routes: A reinforcement learning approach. *Computers & Chemical Engineering* **2020,** *141*.

• A reinforcement learning approach to identify optimal reaction pathways in a reaction network, in contrast to MILP type approaches.

73. Gao, H. Y.; Pauphilet, J.; Struble, T. J.; Coley, C. W.; Jensen, K. F., Direct Optimization across Computer-Generated Reaction Networks Balances Materials Use and Feasibility of Synthesis Plans for Molecule Libraries. *Journal of Chemical Information and Modeling* **2021,** *61* (1), 493-504.

74. Kleinekorte, J.; Kröger, L.; Leonhard, K.; Bardow, A., A neural network-based framework to predict process-specific environmental impacts. In *Computer Aided Chemical Engineering*, Elsevier: 2019; Vol. 46, pp 1447-1452.

75. Kuznetsov, A.; Sahinidis, N. V., ExtractionScore: A Quantitative Framework for Evaluating Synthetic Routes on Predicted Liquid-Liquid Extraction Performance. *Journal of Chemical Information and Modeling* **2021,** *61* (5), 2274-2282.

• Provides a method to compute the ease of separation of major and side products in a synthesis pathway

76. Ryu, J.; Kong, L. X.; de Lima, A. E. P.; Maravelias, C. T., A generalized superstructure-based framework for process synthesis. *Computers & Chemical Engineering* **2020,** *133*.

77. Belotti, P.; Kirches, C.; Leyffer, S.; Linderoth, J.; Luedtke, J.; Mahajan, A., Mixed-integer nonlinear optimization. *Acta Numerica* **2013,** *22*, 1.
78. Floudas, C. A.; Gounaris, C. E., A review of recent advances in global optimization. *Journal of Global Optimization* **2009,** *45* (1), 3-38.
79. Mencarelli, L.; Chen, Q.; Pagot, A.; Grossmann, I. E., A review on superstructure optimization approaches in process system engineering. *Computers & Chemical Engineering* **2020,** *136*.
80. Belotti, P.; Lee, J.; Liberti, L.; Margot, F.; Wachter, A., Branching and bounds tightening techniques for non-convex MINLP. *Optimization Methods & Software* **2009,** *24* (4-5), 597-634.
81. Gottipati, S. K.; Sattarov, B.; Niu, S.; Pathak, Y.; Wei, H.; Liu, S.; Blackburn, S.; Thomas, K.; Coley, C.; Tang, J. In *Learning to navigate the synthetically accessible chemical space using reinforcement learning*, International Conference on Machine Learning, PMLR: 2020; pp 3668-3679.

•• Discuses preliminary efforts to incorporate reinforcement learning to bias forward synthesis

82. Bradshaw, J.; Paige, B.; Kusner, M. J.; Segler, M. H.; Hernández-Lobato, J. M., A model to search for synthesizable molecules. *arXiv preprint arXiv:1906.05221* **2019**.

•• Discusses preliminary efforts to incorporate synthesizeability within generative models